

周报总结

2015-08-23

本周总结：

工作部分：

这一周基本实现了之前 **Evolutionary Clustering** 的论文。之前看论文的时候虽然看懂了，但是动手实现的时候，再来思考一翻又发现了很多问题，感觉这种时序的聚类其实并不十分符合我们目前的可视化方法。

Evolutionary 的聚类方法，核心思路是每一步都要有自己的聚类中心，而当前的聚类中心不是从当前数据得到的，而是要综合之前的聚好的聚类中心，聚出的数据一起得到，并且保证当前的这一类是上一个时间点的同一类演化而来的。具体做法是，初始时间点，用普通的 **kmeans** 来聚出几类，然后之后的时间点，用上一个时刻聚出的中心，综合这一个时间点的的结果，通过一个分配公式，将两个聚类的结果进行综合，得出新的聚类结果。也就是完成了 t 时刻是 $t-1$ 时刻的延续，这种演化的思想。

然而，这种方法并不符合我们之前做的方法的核心思想。我们自己的方法，目标在于：通过聚类的手段，将不同时刻数据分类，通过一种生命线的方式，来展现不同类别之间的转移。所以存在着一种情况：部分类别在时间阶段开始的时候，并不存在，而在阶段中期，才开始产生，增多发展。然而这种 **Evolutionary** 的方法，在开始阶段，就将数据分成了固定的类别，中途产生的类别的现象将不会存在。

同时，由于聚类算法是一种演化的过程，聚类中心，随着时间的变化在缓慢的偏移，这样反而还会弱化我们的转移过程。因为我们的数据也是逐渐变化的，这样在全局聚类的过程中，才会有转移。而演化的方法，反而弱化了这种缓慢变化的过程，因为缓慢的变化很可能被淹没在聚类中心的变化中，这样反而不利于转移的形成。

对于以上两点问题，如果要将这种时序的聚类方法用在我们的论文里，感觉基本的算法方法需要改进。首先，聚类的方法不能是 **off-line** 的（是指当前的聚类不能只与上一阶段的聚类结果有关）而应该是 **online** 的（当前时刻聚类与全局的聚类有关）。如果要对算法进行改进的话，自己的想法是，对于初始聚类，不能简单的使用普通的 **kmeans** 的方法，比如说

随机定中心，而应该是首先对数据进行全局聚类，将生成的全局中心，这样第一步的时候中心使用的是我们全局的中心作为初始值，这样演化开始不是对起始时间的演化，而是对全局的演化。然而这样做可能会存在很多问题，比如说算法的有效性和可用性等，没有做过类似的方向，所以不能进行评估。

因此对于这个方法的使用还是有待商榷的。

未来论文工作：

对于未来的工作，首先要做的是针对之前的评审意见和自己工作中发现的问题进行修改。

1 系统的交互方式比较简单。我们的系统当时由于时间原因，其实很多的交互和细节上做的都十分的粗糙。基本上都是在一种原型的状态。一些交互，比如我们当时的辅助视图的设计，比如说一些基本的拉伸选中，在做的时候都没有进行仔细的思考，尤其是辅助视图的设计，只是用了基本的视图，太过简单，而且功能上还有重合的部分。因此，给人的印象并不是很好。

2 系统的核心方法较为简单。我们的系统基本提供的方法就是将已经聚好的类别展示出来，然后通过点选看看不同类别之间的变化情况。除此之外的能称得上复杂的操作，就是进行了简单的筛选。因此整个系统的功能性太单一。可以考虑增加更多的功能：

1) 巫老师之前说过，对于属性我们可以并不去聚出全部的属性，而是选中其中的几种属性进行聚类。这样的想法其实就可以用在系统之中，这种想法的核心其实是一种更高层次上的筛选，是一种利用了聚类方法之后的筛选。在我们局部聚类之后，我们还可以将局部聚类的结果结合全局聚类的结果进行展示，从而使我们的系统更加丰富。

2) 对于聚类数目的选择，对于 k 的选择我们当时选的是随机选择的，这样也使得几位审稿人都对此提出了意见，因此聚类方法上，比较希望找到一种新的较好的自适应类别数目的方法来做。

3) 可能需要对一些计算过程进行加速优化，尤其是聚类过程，其实我们最希望的是聚类过程尽量实时的完成，这样一些交互才能更加方便的完成。

3 论文的修改，我们的论文当时是本科生撰写的，论文的结构都比较粗糙。整个结构也很混乱，对于文章的大布局，和细节描述而言，因为本科生从来没有写作的经验，因此十分幼稚和混乱。因此，论文的行文和布局可能需要更好的进行斟酌。